Python ではじめる機械学習

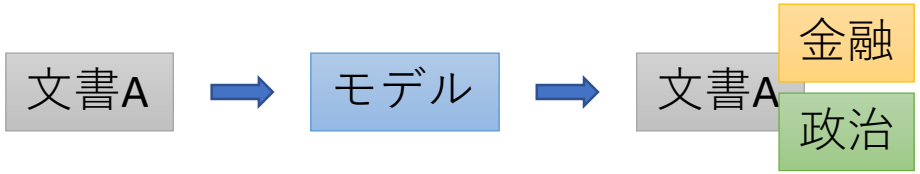
7.9 トピックモデリングと文書クラスタリング

15T4057F　藤井　真

7.9 トピックモデリングと文書クラスタリング

- トピックモデリング（topic modeling）

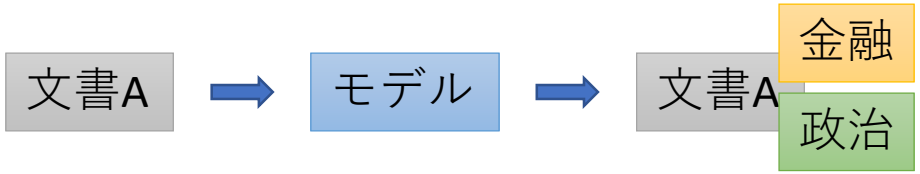
  様々な文書に対して、1つ以上のtopicを割り当てるタスク。
  通常、教師なし学習。



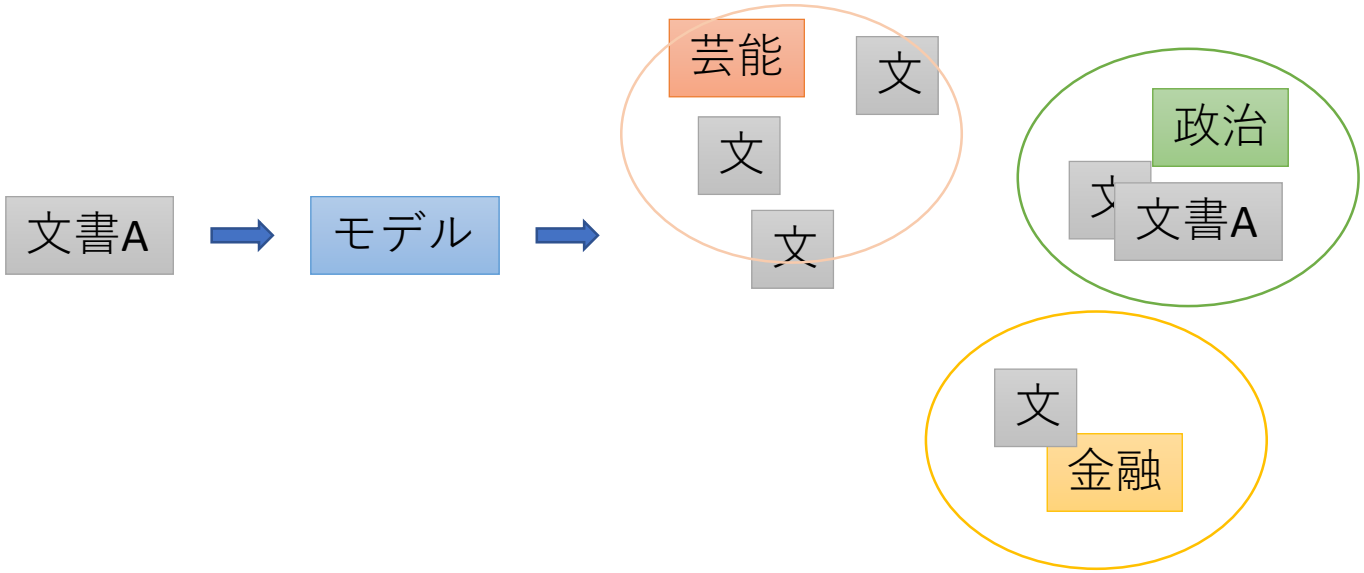
7.9 トピックモデリングと文書クラスタリング

- トピックモデリング（topic modeling）

  様々な文書に対して、1つ以上のtopicを割り当てるタスク。
  通常、教師なし学習。



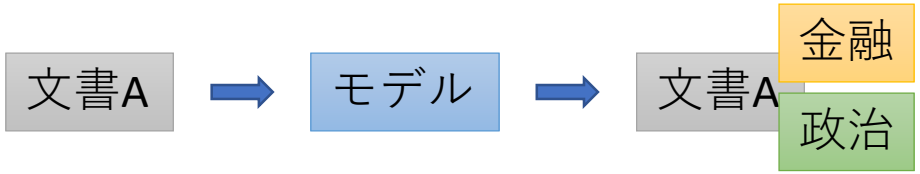
  - １文書に１トピックならクラスタリング的



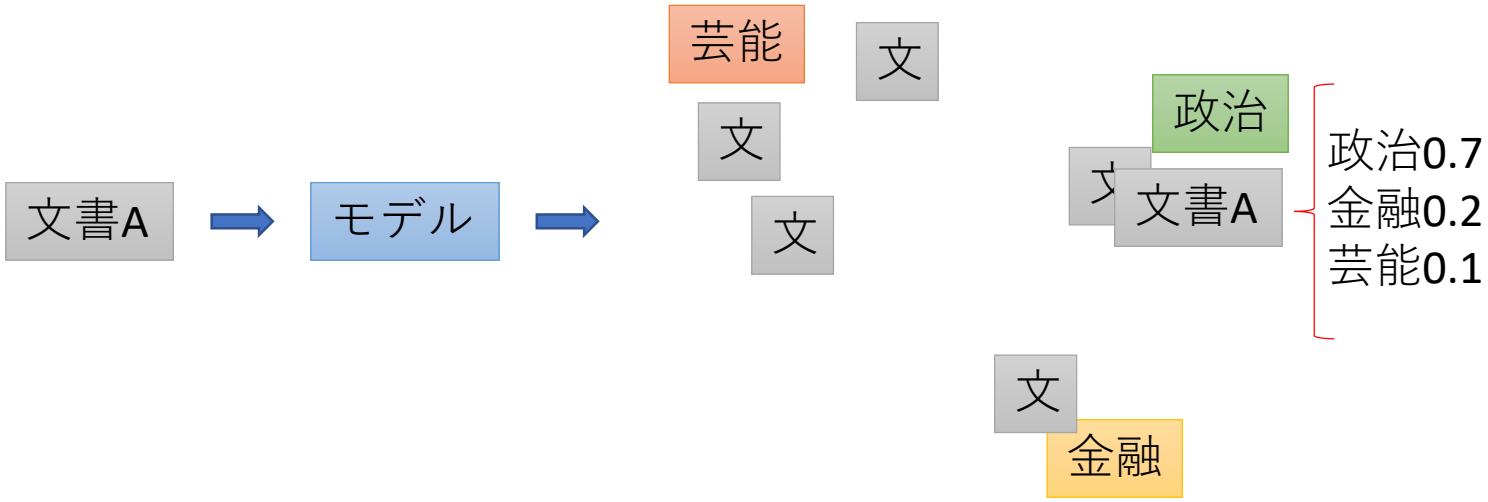
7.9 トピックモデリングと文書クラスタリング

- トピックモデリング（topic modeling）

  様々な文書に対して、1つ以上のtopicを割り当てるタスク。
  通常、教師なし学習。



  - １文書にトピックを2つ以上許せば成分分析的



7.9.1 LDA（Latent Dirichlet Allocation）

隠れている、潜在的 ディリクレ分布 割り当て

- LDA（Latent Dirichlet Allocation）
  - トピックモデリングと言えば、この手法と言える代表。
  - 文章は単語レベルの潜在的トピックの集まり。

    例：BWEとWMDとWSDについて考えてください。・・

7.9.1 LDA（Latent Dirichlet Allocation）

- LDA（Latent Dirichlet Allocation）
  （隠れている、潜在的 ディリクレ分布 割り当て）
  - トピックモデリングと言えば、この手法と言える代表。
  - 文章は単語レベルの潜在的トピックの集まり。

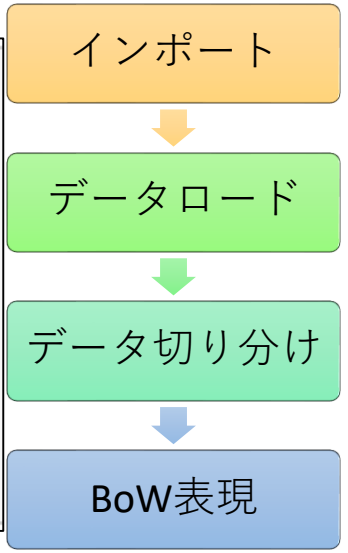
    例：BWEとWMDとWSDについて考えてください。・・

    ⇒各単語からこの文章トピックはNLP
  - 逆に言えば同時に現れやすい単語の集合が文章になる。
  - この同時に現れやすい単語を任意の数に分ければ
    その数の分だけトピックに分けられたと解釈する。

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - データの準備

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.datasets import load_files

reviews_train = load_files("aclImdb/train/")
reviews_test  = load_files("aclImdb/test/")
text_train, y_train = reviews_train.data, reviews_train.target
text_test,  y_test  = reviews_test.data,  reviews_test.target

vect = CountVectorizer(max_features=10000, max_df=.15)
X = vect.fit_transform(text_train)
```



  - 文書の教師なし学習で、一般的な単語は学習を鈍らせることがあるのでBoWモデル作成時
    - 15％以上の文書に登場する単語を除く。
    - それ以外で最頻する10,000単語を用いる。

7.9.1 LDA（Latent Dirichlet Allocation）

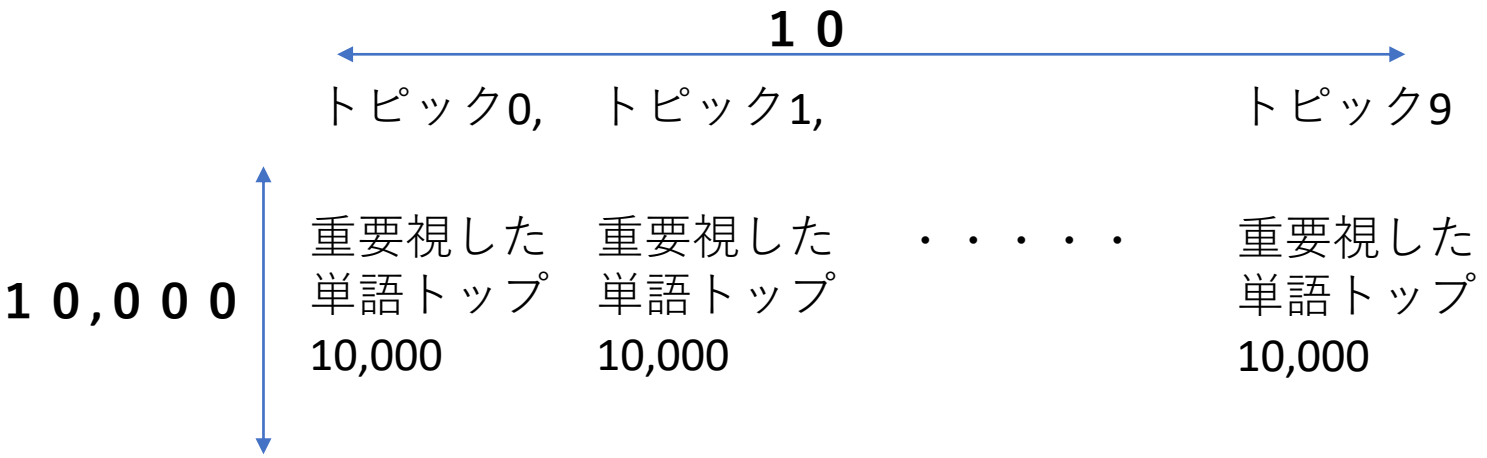
- LDAの映画レビューへの適用
  - LDAモデルの作成

```
from sklearn.decomposition import LatentDirichletAllocation
                                    １０トピック
lda = LatentDirichletAllocation(n_topics=10, learning_method="batch",
                                max_iter=25, random_state=0)
document_topics = lda.fit_transform(X)          結果の固定

print("lda.components_.shape: {}".format(lda.components_.shape))
```

```
lda.components_.shape: (10, 10000)
```

インポート → モデル作成 → print文 → 出力

  - 作成されたモデルイメージ



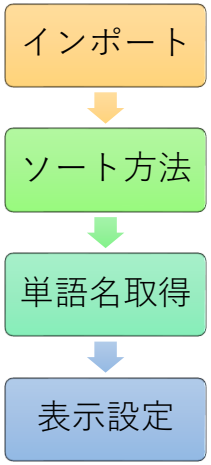
7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １０トピックの重要視した単語１０個を見る。

```
import numpy as np
import mglearn

sorting = np.argsort(lda.components_, axis=1)[:, ::-1]


feature_names = np.array(vect.get_feature_names())

mglearn.tools.print_topics(topics=range(10), feature_names=feature_names,
                           sorting=sorting, topics_per_chunk=5, n_words=10)
```

１０トピック　　特徴量名称

ソート方法　　１０単語



  - print_topics関数を使うと結果をきれいに表示できる。

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １０トピックの重要視した単語１０個を見る。

家系？　コメディ？

topic 0	topic 1	topic 2	topic 3	topic 4
director	show	book	family	funny
work	series	original	young	comedy
performance	war	10	father	cast
actors	episode	now	us	role
cast	tv	again	woman	humor
screen	years	world	own	fun
performances	american	saw	world	jokes
role	episodes	read	real	actors
both	world	didn	mother	performance
quite	shows	am	between	always

ゾンビホラー？　警察系アクション？

topic 5	topic 6	topic 7	topic 8	topic 9
horror	music	original	thing	action
gore	john	team	worst	police
effects	old	series	didn	murder
blood	young	jack	nothing	killer
pretty	girl	action	minutes	crime
budget	song	new	guy	plays
house	gets	down	actually	lee
zombie	dance	tarzan	want	gets
dead	songs	freddy	going	role
low	rock	indian	re	cop

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １０トピックの重要視した単語１０個を見る。

家系？　コメディ？

topic 0	topic 1	topic 2	topic 3	topic 4
director	show	book	family	funny
work	series	original	young	comedy
performance	war	10	father	cast
actors	episode	now	us	role
cast	tv	again	woman	humor
screen	years	world	own	fun
performances	american	saw	world	jokes
role	episodes	read	real	actors
both	world	didn	mother	performance
quite	shows	am	between	always

ゾンビホラー？　警察系アクション？

topic 5	topic 6	topic 7	topic 8	topic 9
horror	music	original	thing	action
gore	john	team	worst	police
effects	old	series	didn	murder
blood	young	jack	nothing	killer
pretty	girl	action	minutes	crime
budget	song	new	guy	plays
house	gets	down	actually	lee
zombie	dance	tarzan	want	gets
dead	songs	freddy	going	role
low	rock	indian	re	cop

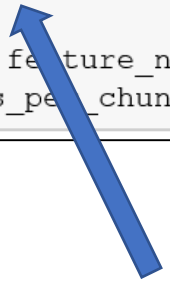
１０トピックで映画全体をカバーするのは、１トピックあたりが幅広くなってぼやける。

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １００トピックなら？

```
lda100 = LatentDirichletAllocation(n_topics=100, learning_method="batch",
                                   max_iter=25, random_state=0)
document_topics100 = lda100.fit_transform(X)
```

```
sorting = np.argsort(lda100.components_, axis=1)[:, ::-1]
feature_names = np.array(vect.get_feature_names())
topics_choiced = np.array([3,4,5,9,23,56,76,78,79,81])
print("\n")
mglearn.tools.print_topics(topics=topics_choiced, feature_names=feature_names,
                           sorting=sorting, topics_per_chunk=5, n_words=10)
```



  - １０トピックとほぼ同様。
  - １００のうち分かりやすい１０トピックを選択。

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １００トピックなら？

キリスト教

topic 3
god
church
christian
jesus
religious
faith
religion
message
christ
believe

バットマン

topic 4
batman
jane
welles
mr
rochester
eyre
timothy
kane
pitt
robin

ゾンビとヴァンプ

topic 5
horror
zombie
gore
zombies
blood
vampire
dead
vampires
pretty
house

マイケルジャクソン

topic 9
music
michael
band
voice
jackson
video
rock
price
fan
clips

日本のアニメ

topic 23
japanese
animation
quite
animated
japan
anime
world
bit
however
english

トムとジェリー

topic 56
tom
jerry
russell
may
anne
kurt
mouse
andrew
night
darren

猟奇

topic 76
killer
girl
killed
death
slasher
killing
kill
gets
horror
dead

家系

topic 78
mother
family
child
daughter
father
son
husband
real
boy
heart

００７

topic 79
peter
allen
woody
bond
james
brosnan
falk
sellers
role
bergman

STAR WARS

topic 81
star
wars
luke
washington
joseph
battle
lucas
smith
denzel
fisher

自分の興味ある
トピックの単語を
確認してみてください。

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １００トピックなら？

キリスト教

topic 3	topic 4	topic 5	topic 9	topic 23
god	batman	horror	music	japanese
church	jane	zombie	michael	animation
christian	welles	gore	band	quite
jesus	mr	zombies	voice	animated
religious	rochester	blood	jackson	japan
faith	eyre	vampire	video	anime
religion	timothy	dead	rock	world
message	kane	vampires	price	bit
christ	pitt	pretty	fan	however
believe	robin	house	clips	english

バットマン

ゾンビと ヴァンプ

マイケル ジャクソン

日本の アニメ

トムと ジェリー

猟奇

家系

００７

STAR WARS

topic 56	topic 76	topic 78	topic 79	topic 81
tom	killer	mother	peter	star
jerry	girl	family	allen	wars
russell	killed	child	woody	luke
may	death	daughter	bond	washington
anne	slasher	father	james	joseph
kurt	killing	son	brosnan	battle
mouse	kill	husband	falk	lucas
andrew	gets	real	sellers	smith
night	horror	boy	role	denzel
darren	dead	heart	bergman	fisher

自分の興味ある
トピックの単語を
確認してみてください。

１０トピックより具体的

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - トピック８１：STARWARS　のレビューを確認する。

```
starwars = np.argsort(document_topics100[:, 81])[::-1]
for i in starwars[:5]:
    print(b".".join(text_train[i].split(b".")[:2]) + b".\n")
```

5文書

2文まで

```
b'Despite the feelings of most "Star Wars" fans, in my opinion "Return Of The Jedi" is the gr
eatest cinematic film ever created. Ever since the first time I saw it, it\'s depth, intensit
y, special effects, and moving story have overwhelmed me.\n'
b'A very good movie. A classic sci-fi film with humor, action and everything.\n'
b"He glorified himself as a great supporting actor in `Glory', he proved he was no `Malcolm i
n the Middle' mediocre actor in `Malcolm X', he showed his brotherly love for acting in `Phil
adelphia', he pulled a slam dunk in `He Got Game', he pulled no punches and rocked us like a
hurricane in `The Hurricane', he provided us effective thespian education in `Training Day',
and now he has demonstrated that he could also direct! Denzel Washington's directorial debut
`Antwone Fisher' is the most moving film of the year. This tearjerker `fish'er story is in no
relation to the debacle that happened to the Miami Dolphins in the 4th quarter against the Ne
w England Patriots in the last game of the 2002 season.\n"
b'Return of the Jedi is certainly the most action packed of the series, and is a fine conclus
ion to the Star Wars Saga. With Han Solo imprisoned by Jabba the Hut and the Empire building
a new Death Star, the rebel alliance is facing an uphill struggle against the dark side, and
only our favourite heroes can pull it off.\n'
b"I noticed that A NEW HOPE and THE EMPIRE STRIKES BACK are in the TOP 10, but that this one
isn't even in the TOP 100.<br /><br />This movie has a bad reputation because of Ewoks, but t
here are so many reasons to love this movie:<br /><br />-The Rescue of Han Solo from Jabba: T
his official wraps up the Han Solo in debt sub-plot that was established when we first met th
e character in A NEW HOPE.\n"
```

３文書目は１文あたりが長く、特殊な内容⇒特徴がとらえずらい

7.9.1 LDA（Latent Dirichlet Allocation）

- LDAの映画レビューへの適用
  - １００トピックが全文書に対して得た重み

```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(1, 2, figsize=(10, 10))
topic_names = ["{:>2} ".format(i) + " ".join(words)
               for i, words in enumerate(feature_names[sorting[:, :2]])]
```

代表的な２単語をトピックタイトルとする

```
for col in [0, 1]:
    start = col * 50
    end = (col + 1) * 50
    ax[col].barh(np.arange(50), np.sum(document_topics100, axis=0)[start:end])
    ax[col].set_yticks(np.arange(50))
    ax[col].set_yticklabels(topic_names[start:end], ha="left", va="top")
    ax[col].invert_yaxis()
    ax[col].set_xlim(0, 2000)
    yax = ax[col].get_yaxis()
    yax.set_tick_params(pad=130)
#plt.tight_layout()
plt.show()
```

インポート

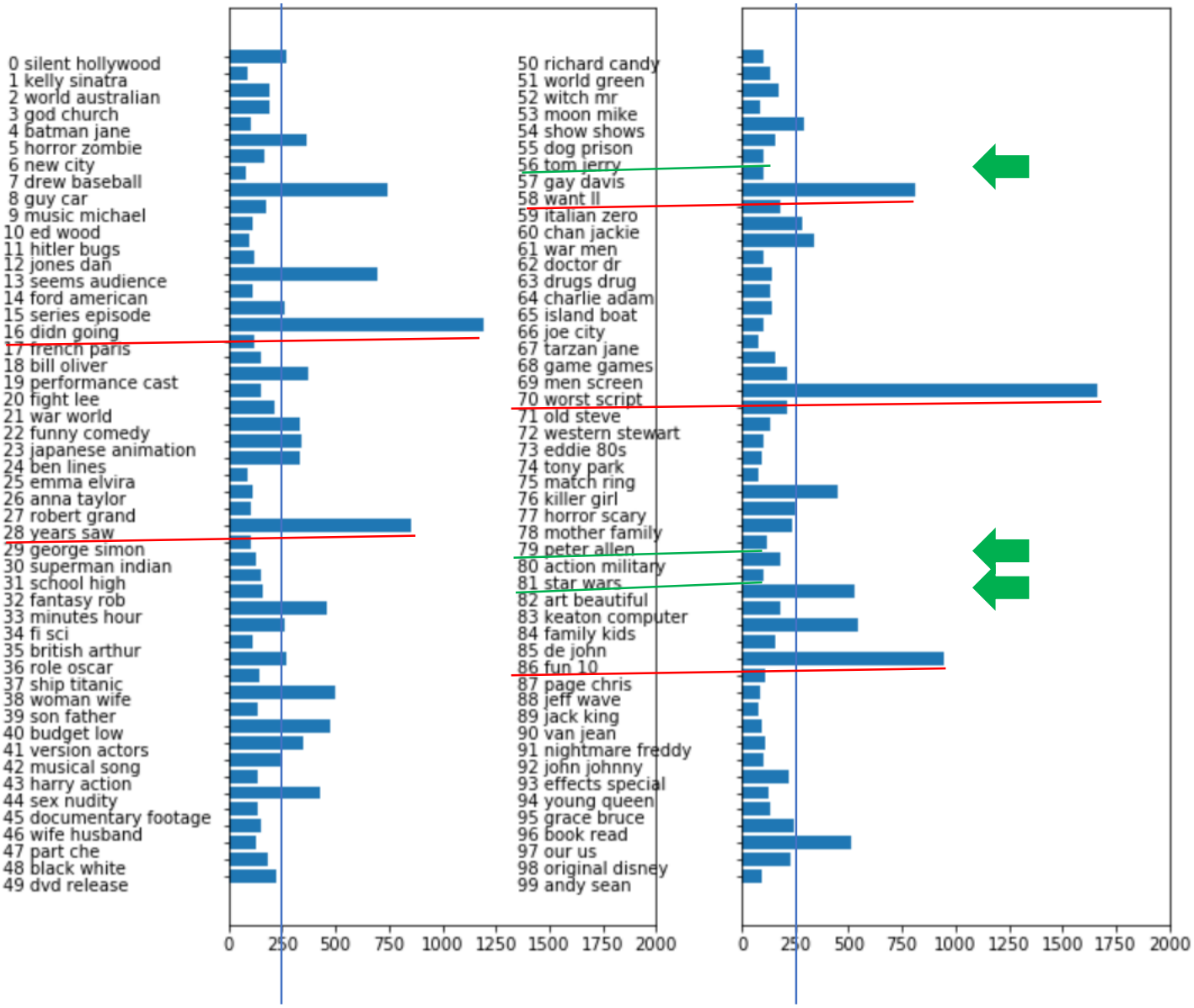
↓

目盛り等
設定

↓

２列で表示
値の取得

7.9.1 LDA（Latent Dirichlet Allocation）



・２５，０００文書
　１００トピック
　平均は２５０文書

・赤線は抽象的な好悪文
　文章を集めやすい

・緑線は上から
　・トムとジェリー
　・００７
　・STARWARS
　具体的で文章が
　集まらない

7.9.1 LDA（Latent Dirichlet Allocation）

- まとめ

  - LDAは「潜在的ディリクレ(分布)割り当て」
  - トピックモデリングを代表する教師なし学習手法
  - トピック数を増やすと具体的トピックが得られる
  - 教師なしらしく人の解釈がいるため
    ただトピック数を増やせばよいわけではない